METHOD FOR SYNTHESIZING SPEECH

FIELD OF THE INVENTION

The present invention relates to the field of analyzing and synthesizing of speech and more particularly without limitation, to the field of text-to-speech synthesis.

BACKGROUND AND PRIOR ART

The function of a text-to-speech (TTS) synthesis system is to synthesize speech from a generic text in a given language. Nowadays, TTS systems have been put into practical operation for many applications, such as access to databases through the telephone network or aid to handicapped people. One method to synthesize speech is by concatenating elements of a recorded set of subunits of speech such as demisyllables or polyphones. The majority of successful commercial systems employ the concatenation of polyphones. The polyphones comprise groups of two (diphones), three (triphones) or more phones and may be determined from nonsense words, by segmenting the desired grouping of phones at stable spectral regions. In a concatenation based synthesis, the conversation of the transition between two adjacent phones is crucial to assure the quality of the synthesized speech. With the choice of polyphones as the basic subunits, the transition between two adjacent phones is preserved in the recorded subunits, and the concatenation is carried out between similar phones.

Before the synthesis, however, the phones must have their duration and pitch modified in order to fulfil the prosodic constraints of the new words containing those phones. This processing is necessary to avoid the production of a monotonous sounding synthesized speech. In a TTS system, this function is performed by a prosodic module. To allow the duration and pitch modifications in the recorded subunits, many concatenation based TTS systems employ the time-domain pitch-synchronous overlap-add (TD-PSOLA) (E. Moulines and F. Charpentier, "Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Commun., vol. 9, pp. 453-467, 1990) model of synthesis.

In the TD-PSOLA model, the speech signal is first submitted to a pitch marking algorithm. This algorithm assigns marks at the peaks of the signal in the voiced segments and assigns marks 10 ms apart in the unvoiced segments. The synthesis is made by a superposition of Hanning windowed segments centered at the pitch marks and extending from the previous pitch mark to the next one. The duration modification is provided by deleting or replicating some of the windowed segments. The pitch period modification, on the other hand, if provided by increasing or decreasing the superposition between windowed segments.

Despite the success achieved in many commercial TTS systems, the synthetic speech produced by using the TD-PSOLA model of synthesis can present some drawbacks, mainly under large prosodic variations, outlined as follows.

1. The pitch modifications introduce a duration modification that needs to be appropriately compensated.
2. The duration modification can only be implemented in a quantized manner, with a one pitch period resolution (α=...,1/2,2/3,3/4,...,4/3,3/2,2/1,...).
3. When performing a duration enlargement in unvoiced portions, the repetition of the segments can introduce "metallic" artifacts (metallic-like sounding of the synthesized speech).

In IEEE transactions on speech and audio processing, vol. 6, No. 5, September 1998, "A Hybrid Model for Text-to-Speech Synthesis", Fábio Violaro and Olivier Böeffard, a hybrid model for concatenation-based, text-to-speech synthesis is described.

The speech signal is submitted to a pitch-synchronous analysis and decomposed into a harmonic component, with a variable maximum frequency, plus a noise component. The harmonic component is modelled as a sum of sinusoids with frequencies multiple of the pitch. The noise component is modelled as a random excitation applied to an LPC filter. In unvoiced segments, the harmonic component is made equal to zero. In the presence of pitch modifications, a new set of harmonic parameters is evaluated by resampling the spectrum envelope at the new harmonic frequencies. For the synthesis of the harmonic component in the presence of duration and / or pitch modifications, a phase correction is introduced into the harmonic parameters.

A variety of other so called "overlap and add" methods are known from the prior art, such as PIOLA (Pitch Inflected OverLap and Add) [P. Meyer, H. W. Rühl, R. Krüger, M. Kugler L.L.M. Vogten, A. Dirksen, and K. Belhoula. PHRITTS: A text-to-speech synthesizer for the German language. In Eurospeech '93, pages 877-890, Berlin, 1993], or PICOLA (Pointer Interval Controlled OverLap and Add) [Morita: "A study on speech expansion and contraction on time axis", Master thesis, Nagoya University (1987), in

Japanese.] These methods differ from each other in the way they mark the pitch period locations.

None of these methods give satisfactory results when applied as a mixer for two different waveforms. The problem is phase mismatches. The phases of harmonics are affected by the recording equipment, room acoustics, distance to the microphone, vowel color, co-articulation effects etc. Some of these factors can be kept unchanged like the recording environment but others like the co-articulation effects are very difficult (if not, impossible) to control. The result is that when pitch period locations are marked without taken into account the phase information, the synthesis quality will suffer from phase mismatches.

Other methods like MBR-PSOLA (Multi Band Resynthesis Pitch Synchronous OverLap Add) [T. Dutoit and H. Leich. MBR-PSOLA: Text-to-speech synthesis based on an MBE re-synthesis of the segments database. Speech Communication, 1993] regenerate the phase information to avoid phase mismatches. But this involves an extra analysis-synthesis operation that reduces the naturalness of the generated speech. The synthesis often sounds mechanic.

US patent 5,787,398 shows an apparatus for synthesizing speech by varying pitch. One of the disadvantages of this approach is that since the pitch marks are centered on the excitation peaks and the measured excitation peak does not necessarily have synchronous phase, phase distortion results.

The pitch of synthesized speech signals is varied by separating the speech signals into a spectral component and an excitation component. The latter is multiplied by a series of overlapping window functions synchronous, in the case of voiced speech, with pitch timing mark information corresponding at least approximately to instants of vocal excitation, to separate it into windowed speech segments which are added together again after the application of a controllable time-shift. The spectral and excitation components are then recombined. The multiplication employs at least two windows per pitch period, each having a duration of less than one pitch period.

US patent 5,081,681 shows a class of methods and related technology for determining the phase of each harmonic from the fundamental frequency of voiced speech. Applications include speech coding, speech enhancement, and time scale modification of speech. The basic approach is to include recreating phase signals from fundamental frequency and voiced/unvoiced information, and adding a random component to the recreated phase signal to improve the quality of the synthesized speech.

US patent No. 5,081,681 describes a method for phase synthesis for speech processing. Since the phase is synthetic the result of the synthesis does not sound natural as many aspects of the human voice and the acoustics of the surround are ignored by the synthesis.

SUMMARY OF THE INVENTION

The present invention provides for a method for analyzing of speech, in particular natural speech. The method for analyzing of speech in accordance with the invention is based on the discovery, that the phase difference between the speech signal, in particular a diphone speech signal, and the first harmonic of the speech signal is a speaker dependent parameter which is basically a constant for different diphones.

In accordance with a preferred embodiment of the invention this phase difference is obtained by determining a maximum of the speech signal and by determining the phase zero, i. e. the positive zero crossing of the first harmonic. The difference between the phases of the maximum and phase zero is the speaker dependent phase difference parameter.

In one application this parameter serves as a basis to determine a window function, such as a raised cosine or a triangular window. Preferably the window function is centered on the phase angle which is given by the zero phase of the first harmonic plus the phase difference. Preferably the window function has its maximum at that phase angle. For example, the window function is chosen to be symmetric with respect to that phase angle.

For speech synthesis diphone samples are windowed by means of the window function, whereby the window function and the diphone sample to be windowed are offset by the phase difference.

The diphone samples which are windowed this way are concatenated. This way the natural phase information is preserved such that the result of the speech synthesis sounds quasi natural.

In accordance with a preferred embodiment of the invention control information is provided which indicates diphones and a pitch contour. For example such control information can be provided by the language processing module of a text-to-speech system.

It is a particular advantage of the present invention in comparison to other time domain overlap and add methods that the pitch period (or the pitch-pulse) locations are synchronized by the phase of the first harmonic.

The phase information can be retrieved by low-pass filtering the first harmonic of the original speech signal and using the positive zero-crossing as indicators of zero-phase. This way, the phase discontinuity artefacts are avoided without changing the original phase information.

Applications for the speech synthesis methods and the speech synthesis device of the invention include: telecommunication services, language education, aid to handicapped persons, talking books and toys, vocal monitoring, multimedia, man-machine communication.

BRIEF DESCRIPTION OF THE DRAWINGS

In the following preferred embodiments of the invention are described in greater detail by making reference to the drawings in which :

Figure 1 is illustrative of a flow chart of a method to determine the phase difference between a diphone at its first harmonic,

Figure 2 is illustrative of signal diagrams to illustrate an example of the application of the method of Figure 1,

Figure 3 is illustrative of an embodiment of the method of the invention for synthesizing speech,

Figure 4 shows an application example of the method of Figure 3,

Figure 5 is illustrative of an application of the invention for processing of natural speech,

Figure 6 is illustrative of an application of the invention for text-to-speech,

Figure 7 is an example of a file containing phonetic information,

Figure 8 is an example of a file containing diphone information extracted from the file of Figure 7,

Figure 9 is illustrative of the result of a processing of the files of Figures 7 and 8,

Figure 10 shows a block diagram of a speech analysis and synthesis apparatus in accordance with the present invention.

DETAILED DESCRIPTION

The flow chart of Figure 1 is illustrative of a method for speech analysis in accordance with the present invention. In step 101 natural speech is inputted. For the input of natural speech known training sequences of nonsense words can be utilized. In step 102 diphones are extracted from the natural speech. The diphones are cut from the natural speech and consist of the transition from one phoneme to the other.

In the next step 103 at least one of the diphones is low-pass filtered to obtain the first harmonic of the diphone. This first harmonic is a speaker dependent characteristic which can be kept constant during the recordings.

In step 104 the phase difference between the first harmonic and the diphone is determined. Again this phase difference is a speaker specific voice parameter. This parameter is useful for speech synthesis as will be explained in more detail with respect to Figures 3 to 10.

Figure 2 is illustrative of one method to determine the phase difference between the first harmonic and the diphone (cf. step 4 of Figure 1). A sound wave 201 acquired from natural speech forms the basis for the analysis. The sound wave 201 is low-pass filtered with a cut-off frequency of about 150 Hz in order to obtain the first harmonic 202 of the sound wave 201. The positive zero-crossings of the first harmonic 202 define the phase angle zero. The first harmonic 202 as depicted in Figure 2 covers a number of 19 succeeding complete periods. In the example considered here the duration of the periods slightly increases from period 1 to period 19. For one of the periods the local maximum of the sound waveform 201 within that period is determined.

For example the local maximum of the sound wave 201 within the period 1 is the maximum 203. The phase of the maximum 203 within the period 1 is denoted as φ_{max} in Figure 2. The difference $\Delta\varphi$ between φ_{max} and the zero phase φ_0 of the period 1 is a speaker dependent speech parameter. In the example considered here this phase difference is about 0,3 π. It is to be noted that this phase difference is about constant irrespective of which one of the maxima is utilized in order to determine this phase difference. It is however preferable to choose a period with a distinctive maximum energy location for this measurement. For example if the maximum 204 within the period 9 is utilized to perform this analysis the resulting phase difference is about the same as for the period 1.

Figure 3 is illustrative of an application of the speech synthesis method of the invention. In step 301 diphones which have been obtained from natural speech are windowed by a window function which has its maximum at $\varphi_0+\Delta\varphi$; for example a raised cosine which is centered with respect to the phase $\varphi_0+\Delta\varphi$ can be chosen.

This way pitch bells of the diphones are provided in step 302. In step 303 speech information is inputted. This can be information which has been obtained from natural speech or from a text-to-speech system, such as the language processing module of such a text-to-speech system.

In accordance with the speech information pitch bells are selected. For instance the speech information contains information of the diphones and of the pitch contour to be synthesized. In this case the pitch bells are selected accordingly in step 304 such that the concatenation of the pitch bells in step 305 results in the desired speech output in step 306.

An application of the method of Figure 3 is illustrated by way of example in Figure 4. Figure 4 shows a sound wave 401 which consists of a number of diphones. The analysis as explained with respect to Figures 1 and 2 above is applied to the sound wave 401 in order to obtain the zero phase φ_0 for each of the pitch intervals. As in the example of Figure 2 the zero phase φ_0 is offset from the phase φ_{max} of the maximum within the pitch interval by a phase angle of $\Delta\varphi$ which is about constant.

A raised cosine 402 is used to window the sound wave 401. The raised cosine 402 is centered with respect to the phase $\varphi_0+\Delta\varphi$. Windowing of the sound wave 401 by means of the raised cosine 402 provides successive pitch bells 403. This way the diphone waveforms of the sound wave 401 are split into such successive pitch bells 403. The pitch bells 403 are obtained from two neighboring periods by means of the raised cosine which is centered to the phase $\varphi_0+\Delta\varphi$. An advantage of utilizing a raised cosine rather than a rectangular function is that the edges are smooth this way. It is to be noted that this operation is reversible by overlapping and adding all of the pitch bells 403 in the same order; this produces about the original sound wave 401.

The duration of the sound wave 401 can be changed by repeating or skipping pitch bells 403 and / or by moving the pitch bells 403 towards or from each other in order to change the pitch. The sound wave 404 is synthesized this way by repeating the same pitch bell 403 with a higher than the original pitch in order to increase the original pitch of the sound wave 401. It is to be noted that the phases remain in tact as a result of this overlapping operation because of the prior window operation which has been performed taking into account the characteristic phase difference Δφ. This way pitch bells 403 can be utilized as building blocks in order to synthesize quasi-natural speech.

Figure 5 illustrates one application for processing of natural speech. In step 501 natural speech of a known speaker is inputted. This corresponds to inputting of a sound wave 401 as depicted in Figure 4. The natural speech is windowed by the raised cosine 402 (cf. Figure 4) or by another suitable window function which is centered with respect to the zero phase $\varphi_0+\Delta\varphi$.

This way the natural speech is decomposed into pitch bells (cf. pitch bell 403 of Figure 4) which are provided in step 503.

In step 504 the pitch bells provided in step 503 are utilized as "building blocks" for speech synthesis. One way of processing is to leave the pitch bells as such unchanged but leave out certain pitch bells or to repeat certain pitch bells. For example if every fourth pitch bell is left out this increases the speed of the speech by 25 % without otherwise altering the sound of the speech. Likewise the speech speed can be decreased by repeating certain pitch bells.

Alternatively or in addition the distance of the pitch bells is modified in order to increase or decrease the pitch.

In step 505 the processed pitch bells are overlapped in order to produce a synthetic speech waveform which sounds quasi natural.

Figure 6 is illustrative of another application of the present invention. In step 601 speech information is provided. The speech information comprises phonemes, duration of the phonemes and pitch information. Such speech information can be generated from text by a state of the art text-to-speech processing system.

From this speech information provided in step 601 the diphones are extracted in step 602. In step 603 the required diphone locations on the time axis and the pitch contour is determined based on the information provided in step 601.

In step 604 pitch bells are selected in accordance with the timing and pitch requirements as determined in step 603. The selected pitch bells are concatenated to provide a quasi natural speech output in step 605.

This procedure is further illustrated by means of an example as shown in Figures 7 to 9.

Figure 7 shows a phonetic transcription of the sentence "HELLO WORLD!". The first column 701 of the transcription contains the phonemes in the SAMPA standard notation. The second column 702 indicates the duration of the individual phonemes in milliseconds. The third column comprises pitch information. A pitch movement is denoted by two numbers: position, as a percentage of the phoneme duration, and the pitch frequency in Hz.

The synthesis starts with the search in a previously generated database of diphones. The diphones are cut from real speech and consist of the transition from one phoneme to the other. All possible phoneme combinations for a certain language have to be stored in this database along with some extra information like the phoneme boundary. If there are multiple databases of different speakers, the choice of a certain speaker can be an extra input to the synthesizer.

Figure 8 shows the diphones for the sentence "HELLO WORLD!", i.e. all phoneme transitions in the column 701 of Figure 7.

Figure 9 shows the result of a calculation of the location of the phoneme boundaries, diphone boundaries and pitch period locations which are to be synthesized. The phoneme boundaries are calculated by adding the phoneme durations. For example the phoneme "h" starts after 100 ms of silence. The phoneme "schwa" starts after 155 ms = 100 ms + 55 ms, and so on.

The diphone boundaries are retrieved from the database as a percentage of the phoneme duration. Both the location of the individual phonemes as well as the diphone boundaries are indicated in the upper diagram 901 in Figure 9, where the starting points of the diphones are indicated. The starting points are calculated based on the phoneme duration given by column 702 and the percentage of phoneme duration given in column 703.

The diagram 902 of Figure 9 shows the pitch contour of "HELLO WORLD!". The pitch contour is determined based on the pitch information contained in the column 703 (cf. Figure 7). For example, if the current pitch location is at 0,25 seconds than the pitch period would be at 50 % of the first 'l' phoneme. The corresponding pitch lies between 133 and 139 Hz. It can be calculated with a linear equation:

$$\frac{(0.8\cdot 63+0.5\cdot 64)\cdot 133+(0.2\cdot 128+0.5\cdot 64)\cdot 139}{0.8\cdot 63+64+0.2\cdot 128}=135.5Hz \qquad (1)$$

The next pitch location would than be at 0.2500 + 1/135,5 = 0.2574 seconds. It is also possible to use a non-linear function (like the ERB-rate scale) for this calculation. The ERB (equivalent rectangular bandwidth) is a scale that is derived from psycho-acoustic measurements (Glasberg and Moore, 1990) and gives a better representation by taking into account the masking properties of the human ear. The formula for the frequency to ERB-transformation is:

$$ERB(f) = 21.4 \cdot \log^{10}(4.37 \cdot f) \quad (2)$$

where f is the frequency in kHz. Thc idea is that the pitch changes in the ERB-rate scale are perceived by the human ear as linear changes.

Note that unvoiced regions are also marked with pitch period locations even though unvoiced parts have no pitch.

The varying pitch is given by the pitch contour in the diagram 902 is also illustrated within the diagram 901 by means of the vertical lines 903 which have varying distances. The greater the distance between two lines 903 the lower the pitch. The phoneme, diphone and pitch information given in the diagrams 901 and 902 is the specification for the speech to be synthesized. Diphone samples, i.e. pitch bells (cf. pitch bell 403 of Figure 4) are taken from a diphone database. For each of the diphones a number of such pitch bells for that diphone is concatenated with a number of pitch bells corresponding to the duration of the diphone and a distance between the pitch bells corresponding to the required pitch frequency as given by the pitch contour in the diagram of 902.

The result of the concatenation of all pitch bells is a quasi natural synthesized speech. This is because phase related discontinuities at diphone boundaries are prevented by means of the present invention. This compares to the prior art where such discontinuities are unavoidable due to phase mismatches of the pitch periods.

Also the prosody (pitch /duration) is correct, as the duration of both sides of each diphone has been correctly adjusted. Also the pitch matches the desired pitch contour function.

Figure 10 shows an apparatus 950, such as a personal computer, which has been programmed to implement the present invention. The apparatus 950 has a speech analysis module 951 which serves to determine the characteristic phase difference $\Delta\varphi$. For this purpose the speech analysis module 951 has a storage 952 in order to store one diphone speech wave. In order to obtain the constant phase difference $\Delta\varphi$ only one diphone is sufficient.

Further the speech analysis module 951 has a low-pass filter module 953. The low-pass filter module 953 has a cut-off frequency of about 150 Hz, or another suitable cut-off frequency, in order to filter out the first harmonic of the diphone stored in the storage 952.

The module 954 of the apparatus 950 serves to determine the distance between a maximum energy location within a certain period of the diphone and its first harmonic zero phase location (this distance is transformed into the phase difference Δφ). This can be done by determining the phase difference between zero phase as given by the positive zero crossing of the first harmonic and the maximum of the diphone within that period of the harmonic as it has been illustrated in the example of Figure 2.

As a result of the speech analysis the speech analysis module 951 provides the characteristic phase difference Δφ and thus for all the diphones in the database the period locations (on which e.g. the raised cosine windows are centered to get the pitch-bells). The phase difference Δφ is stored in storage 955.

The apparatus 950 further has a speech synthesis module 956. The speech synthesis module 956 has storage 957 for storing of pitch bells, i.e. diphone samples which have been windowed by means of the window function as it is also illustrated in Figure 2. It is to be noted that the storage 957 does not necessarily have to be bitch-bells. The whole diphones can be stored with period location information, or the diphones can be monotonized to a constant pitch. This way it is possible to retrieve bitch-bells from the database by using a window function in the synthesis module.

The module 958 serves to select pitch bells and to adapt the pitch bells to the required pitch. This is done based on control information provided to the module 958.

The module 959 serves to concatenate the pitch bells selected in the module 958 to provide a speech output by means of module 960.

List of reference numerals

sound wave	201
first harmonic	202
maximum	203
maximum	204
sound wave	401
raised cosine	402
pitch bell	403
sound wave	404
column	701
column	702
column	703
diagram	901
diagram	902
apparatus	950
speech analysis module	951
storage	952
low pass filter module	953
module	954
storage	955
speech synthesis module	956
storage	957
module	958
module	959
module	960